CHAPTER 4

Health needs and their assessment: demography and epidemiology

Chapter contents

Introduction

The relationship between public health and the assessment of need for health care can be traced back to the Acheson Report in the UK (Acheson 1988), which called for regular reviews of the nation's health, and health service reforms formalised this with splits between purchasers and providers of health care (Secretaries of State for

Health, Wales, Northern Ireland and Scotland 1989a, 1989b). The effect has been to place multidisciplinary research on health and effectiveness of health services, the development of evidence-based health care and the assessment of need for services firmly on international agendas of research and development (Peckham 1991; Department of Health 1993a). Health needs assessment is a tool which is now used internationally, including Europe and the USA. Part 4.1 of this chapter focuses on the concept of need and the practice of measuring needs for health services. Two main disciplines in this area are epidemiology and demography; their principal methods and techniques are described in Parts 4.2 and 4.3.

4.1 The assessment of health needs

Health needs

Health needs assessments have traditionally been undertaken by public health professionals in relation to their local populations, though the approach is relevant to all health care sectors, in all populations, and in all countries. As governments across Europe are also faced with rising demands for health care, limited resources and increasing inequalities in health, the World Health Organization (WHO) European Office published a tool describing the stages of community health needs assessment, at the level of families, communities and populations. It was aimed at family health nurses (nurses, midwives and public health nurses) (Rowe *et al.* 2001), to enable practitioners, managers and policy-makers to identify priority health needs, target health care resources to address inequalities, and involve local people, specifically in doing the following:

- to plan and deliver the most effective care to those in greatest need;
- to apply principles of equity and social justice in practice;
- to ensure that scarce resources are allocated where they give maximum health benefit;
- to work collaboratively with the community, other professionals and agencies to determine and prioritise the health issues that cause greatest concern and plan interventions to address them.

The assessment of health needs is a contentious area, and considerable confusion exists about the meaning of 'needs' (Frankel 1991). This stems from the different imperatives that influence the relationship between 'needs' and the provision of health care. The *public health imperative* is concerned with total population needs and the development of strategies based on prevention and health promotion. The *economic imperative* is concerned with marginal met needs and the most efficient ways of meeting needs. The *political imperative* has been one of reconciling a welfare system to the demands of the free market ideology (Jones 1995). The relationship between needs and welfare provision has received considerable critical attention, with the debate focusing on *absolute, normative* and *relative* definitions of need (Soper 1981; Wiggins and Dermen 1987; Doyal and Gough 1991).

Thus, there are multiple perspectives of need to incorporate: perceived and expressed needs of the profiled population; perceptions of professionals providing the services; perceptions and priorities of managers of commissioner/provider organisations (regarding

national, regional or local priorities). Health needs assessment must balance these different perspectives when identifying priorities, and negiotiate with stakeholders to effect the changes required to improve health and reduce health inequalities.

Rowe *et al.* (2001), in their WHO guide to needs assessment, use a holistic model of health, emphasising the social, economic and cultural factors that affect health, plus individual behaviour. The concept of 'need' used incorporates those needs felt and expressed by local people as well as those defined by professionals. It moves beyond the concept of demand, and includes people's capacity to benefit from health care and public health initiatives.

However, the question of whether health needs assessment has actually helped to improve health or reduce inequalities, or whether changes would have happened anyway, remains unanswered.

The need for health and the need for health care

It is important to distinguish between the need for health and the need for health care. Health care is one way of satisfying the need for health. Arguments in the past have concentrated on the relationships between needs and the demand for, access to and use of services (Last 1963; Titmuss 1968; Hart 1971). In this sense, need is not an absolute concept, but is relative and dependent on socio-economic and cultural factors as well as supply-side factors. The need for health was perceived by Acheson (1978) as relief from the negative states of distress, discomfort, disability, handicap and the risk of mortality and morbidity. These concepts form the basis of, but do not wholly determine the need for, health services. This amounts to a bio-medical approach to health care needs that lends itself to the quantitative measurement of health status; the resulting health care needs reported fit conveniently with the bio-medical focus on the incidence and prevalence of disease.

Bradshaw (1972), on the other hand, constructed a paradigm of need in terms of: *expressed need* ('demand'), which is the expression in action of felt need; *comparative need*, which involves comparisons with the situation of others and considerations of equity; and *normative need*, such as experts' definitions, that change over time in response to knowledge. The expressions of need using these definitions are not necessarily consistent in relation to any individual. For many conditions, perceived need for care depends on the beliefs and knowledge of the person affected, and hence on value judgements (Buchan *et al.* 1990). In turn, these are influenced by psychological, socio-economic and cultural factors, not simply by the supply of services. Bradshaw (1994) later acknowledged the weaknesses of his original classification of need, but argued that it was never intended to form a hierarchy of needs. However, his paradigm forms a sociological approach that sets up a useful definitional matrix for needs.

Economists have consistently argued against the concept of objective need (Culyer 1995), seeing need as relative but at the same time recognising its practical importance and proposing concepts such as marginal met needs or, in relation to health care, the capacity to benefit from treatment. For example, Buchan *et al.* (1990) defined need as follows: 'People in need of a health service are defined as those for whom an intervention produces a benefit at reasonable risk and acceptable cost.' Culyer and Wagstaff (1991) considered the relationship between economic evaluation and need in detail, and proffered a precise definition of need that relates specifically to health care:

'A need for medical care is then said to exist so long as the marginal product of care is positive, i.e. so long as the individual's capacity to benefit from medical care is positive.' Economists have also emphasised the importance of health service priorities, given the scarcity of societal resources (Williams 1992). The debate has prompted some to argue that health care needs cannot be discussed in isolation from other needs (Seedhouse 1994), though in Britain, while national NHS policy recognises the importance of the views of the public in defining needs, there is less interest in the latter at local level, partly because health authorities do not know what to do with the results if they cannot clearly relate them to the need for effective services. As Fitzpatrick (1994) put it, 'From the health care provider's perspective, subjective health status problems are insufficiently specific to identify levels of medically determined need for particular health care interventions.'

Doyal and Gough (1991) constructed a theory of human needs based on the notion of basic needs being health and autonomy, an optimum level of which is fundamental to allow participation in social life. Thus, health care becomes a means of satisfying basic need. Soper (1993) sympathises with their argument but contests that their theory collapses when it is applied to specific needs. It is with this problematic specific level that health services researchers and planners have to deal. The orthodox response seems to be to follow the economic line and define needs in relation to supply. What is clear, however, is that if the meeting of needs is to be democratic, then they have to be debated openly. This means democratising the process of needs assessment so that individuals and communities are able to participate fully in decision-making about services. Such participation should extend beyond opinion polls and surveys to involvement in research and needs assessment itself.

Need for effective health care

Data from consumer consultation exercises, health surveys, mortality and morbidity statistics, and other information on the 'need for health' do not indicate to health planners what can be done to improve health (Stevens 1991). Thus, health planners prefer to base health need on a disease model and define it in relation to the need for *effective* health care and preventive services. Although a subsequent document produced by the NHS Management Executive (1991), and documents that followed it, modified this definition to include taking the views of interested parties into account in order to develop an overall understanding of need, and to be responsive to the views of local people about the patterns and delivery of services, the narrower definition has become that most widely used by health planners and public health specialists. Using this definition, need is linked to the appropriateness and effectiveness of the intervention in question. There is, however, considerable uncertainty about the appropriateness of different treatments, as reflected in variations in medical and surgical practice (Evans 1990). Any attempt to define health care needs is always open to criticisms of having a dual role of subjugating the individual or group being assessed to the needs of the system or professional interests within the system, while simultaneously constructing a picture of what that individual or group 'needs' (Jones 1995).

Social variations

While it is arguable that a health service agenda cannot take on the wider definition of need, which is affected by the social structure of a society, it should be concerned with tackling

variations in health care provision to ensure equity, as well as understanding the contribution services can make to mitigating social variations in health, which are also related to the distribution of income and the degree of inequality in society (see Bradshaw 1994). As Popay and Williams (1994) stated, lay knowledge about health, illness and care is vital for understanding the experience of ill health and the processes and outcomes of health and social care. They pointed to 'the need to take seriously people's own views about their health and their health needs', which traditional epidemiological techniques are unable to make accessible, and to the increasing importance of the role of social scientists in research on people's health. Fitzpatrick (1994) also argued that the epidemiological techniques of documenting incidence and prevalence of illnesses and chronic conditions are not the same as identifying needs for health care. The issue of service effectiveness apart, he points to the vital role of the social sciences in developing an understanding of the patient's perspective regarding his or her illness, which should sensitise health professionals to his or her needs. The role of social science was described further in Chapter 2.

Local engagement

Some purchasers of health care do attempt to involve local people in the planning process by holding focus group meetings, or conducting surveys of their views and concerns, their health and their views for health priorities. Some undertake action research or rapid appraisal projects in local communities to achieve this end. Ong and Humphris (1994) argued that needs assessment requires a multidisciplinary approach and that

> The expertise held by users and communities has to be an integral part of needs assessment and to be considered alongside the public-health and clinical-needs assessments. The different inputs in the needs-assessment process offer specific and complementary insights on the complexity of needs as experienced by individuals and populations.

They recommend methods that combine a community perspective and a dialogue with decision-makers (e.g. rapid appraisal). Such techniques must be seen within a larger programme of the assessment of health needs, because they focus on felt and expressed need, rather than epidemiological or clinical assessments of need.

The narrow definition of health need as need for effective services also underpins the contracting process in health services. The underlying philosophy of this conception of need is related to prioritisation of health services and health service rationing, given that health needs are infinite and health care resources are limited. Ideal practice is to maximise the total amount of benefit within existing resources. This raises the problem of finding a method of prioritising health services, which is still unresolved (Bowling 1996a), though the QALY – or quality-adjusted life year – underpins treatments approved by some organisations (e.g. the National Institute for Health and Clinical Excellence, which is the independent organisation responsible for providing national guidance in England and Wales on health care, and the cost-effectiveness of new health technologies) (NICE 2004; http://www.nice.org.uk/pdf/CG011fullguideline.pdf; accessed September 2013).

The health services' research definition of need also makes the assumption that needs can only be met by a health service where adequate information exists about the cost-effectiveness of services. This has led to an active international research industry in systematic reviews and in health technology assessment (http://www.ncchta.org) (Oxman 1996).

Methods of assessing health needs

Cavanagh and Chadwick (2005) described health needs assessment as:

- a systematic method for reviewing the health issues facing a population – in an *ideal* world leading to agreed priorities and resource allocation to demonstrably improve health outcomes and reduce inequalities;
- a recommended public health tool for reviewing health issues;
- an opportunity to engage with specific populations;
- an opportunity for cross-sector partnership working.

A step-by-step guide to conducting needs assessment was produced by Cavanagh and Chadwick (2005). Rowe *et al.* (2001), in their WHO guide for nurses in Europe, also described the steps of community health needs assessment, referring to it as a developmental process, added to and amended over time. It describes the state of health of local people; enables the identification of the major risk factors and causes of ill health; and enables the identification of the actions needed to address these. The steps of community health needs assessment were listed as:

- profiling: the collection of relevant information about the state of health and health needs of the population;
- analysing this information to identify the major health issues;
- deciding on priorities for action;
- planning public health and health care to address the priority issues;
- implementing the planned activities;
- evaluation of health outcomes.

The authors emphasised the need to work in partnership with local people and collaborate with other professionals.

Information to collect

The measurement of need requires information about the level of morbidity (i.e. the size of the health problem) in a given population, the burden on that population and the impact the intervention is likely to have. The information required to address this includes data about the different types of treatments and services that are available in relation to the condition, their effectiveness and cost-effectiveness. This also raises the issue of how to measure burden and effectiveness (see Table 4.1).

The first decision to be made when assessing needs for health services in a particular area is which condition to start with. This will be influenced by local priorities, which in turn are influenced by mortality patterns and standardised mortality ratios (SMRs) in the area. For example, if there is a high rate of coronary heart disease, and a higher mortality rate than the adjusted average (as measured by the SMR), then this may be considered as a priority for action.

The range of techniques includes: calculation of existing health service activity levels and resource norms; calculation of rates of clinical procedures and treatments for specific conditions by population group; estimation of the prevalence of disease in the population and determination of appropriate intervention levels (i.e. the number in the population with a given set of indications for treatment); and application of social deprivation indicators to populations where social deprivation influences need. For some procedures, and in

Theory	Practice	Gap
Agree the disease/condition for assessment and the diagnostic categories to be used.	Agree the disease/condition for assessment and the diagnostic categories to be used.	Medicalisation of needs. Definitions contested (e.g. disability, mental illness).
Define the population served.	Define the population served.	Populations are not static. Who is counted? Who is excluded? (e.g. non-random census undercounting).
Identify the range of treatments and services provided locally and elsewhere.	Identify the range of treatments and services provided locally.	
	Review the literature on incidence and prevalence of the disease, risk factors, mortality rates, the range of treatments and services offered, their effectiveness, cost-effectiveness and levels of appropriateness.	Burgeoning literature. Problems of meta-analysis. Importance of Cochrane reviews and databases.
Establish criteria of appropriateness for the health service intervention.	Apply this knowledge to the population of interest, taking local information into account.	
Establish the effectiveness and costs of each treatment and service.		Problems in obtaining accurate, reliable and comparable costing data.
Estimate the numbers in the target population, the numbers in the diagnostic group selected in that population, and the numbers likely to benefit from each type of intervention.	Build up neighbourhood profiles on health, mortality, socio-demographic characteristics, available services, access to, and use of, services. Local health surveys and rapid appraisal techniques, which involve the public and key professionals, might be used. Match this data, along with demographic and epidemiological disease profiles, to service availability.	Limitations of census data for particular populations. Local data sources (e.g. registers) may lack coverage. Routine data sources may be incomplete. Health surveys: expensive sampling problems response problems translation?

	Undertake comparative assessments of service type and level between districts. Identify gaps in routine information and research with a view to carrying out an epidemiological survey or apply data from elsewhere. Identify the strengths and weaknesses of providers (e.g. waiting lists, referral patterns, treatment delays, intervention rates, rehabilitation and prevention procedures); compare with providers of health care elsewhere.	Rapid appraisal: robust? reliable? generalisable? Selection of control districts. Norms-based approach. Problems of league tables and controlling for case-mix.
Set standards for monitoring and the level of resources required for effective provision of care.	Establish programmes to evaluate the outcome of services and treatments, and their costs, where existing information is inadequate, and calculate the proportion of people with the condition who would benefit from their supply. Establish mechanisms with clinicians at local levels to agree on thresholds for treatment and monitoring of contracts. Monitor the impact of health service contracts with providers in relation to the needs of the population (e.g. number on the waiting list for a specific procedure, number of procedures performed).	Outcome data limited. Often not collected routinely. Consensus panel work difficult, local autonomy may be strong. Professional resistance. Ownership of data may be a problem related to the tension between cooperation and market imperatives. Exclusion.
Community participation at all levels of needs assessment process.	Include the expertise of the public as (potential) users of health services (e.g. through rapid appraisal methods).	Public apathy. Barriers. Professional frustrations. Lack of accountability. Ethical and political objections. Democratic deficit.

Table 4.1 Assessment of health needs: comparison of ideal with practice

certain areas, an adjustment might need to be made for the proportion of the population absorbed by the private health sector. The assessment of health needs, then, involves a combination of the epidemiological assessment of disease prevalence, the evaluation of the effectiveness of treatment and care options, and their relative costs and effectiveness, analysis of existing activity and resource data, and the application of this knowledge to populations (in the case of health authorities to local populations, and in the case of general practitioners to their practice populations or catchment areas). It should also include the expertise of the public as (potential) users of health services (e.g. through rapid appraisal methods).

Because epidemiological surveys are expensive and time-consuming, one alternative is to apply the prevalence ratios and incidence rates reported in the literature to the population targeted (Purcell and Kish 1979; Wilcock 1979; Mackenzie *et al.* 1985). In some areas (e.g. heterogeneous inner-city populations), the level of inaccuracy with this approach will be too high to be acceptable. For example, variations in socio-economic group and the ethnic status of the population can affect the applicability of national data, or data from other areas, to local situations.

One approach that has been suggested is to compare existing service levels with those expected from the population covered, and to investigate further any specialties showing an unexpectedly high or low utilisation rate (Kirkup and Forster 1990). In some cases, it is certainly possible to compare existing service provision in districts with the number of cases that would be expected if national utilisation rates were applied. However, this is unlikely to lead to accurate estimates of need given that service use is affected by so many variables (e.g. resource allocation and supply, historical and political factors, and the patients' perceptions of health and level of knowledge).

In practice, it is unlikely that the information to do this will be available. The information that is available and currently used by health districts (departments of public health) to assess health needs in Britain falls short of the true epidemiological assessment of needs (Stevens 1991). Nevertheless, the information available includes national demographic statistics on mortality and fertility, small area statistics from census data and other sources on the social characteristics of areas which are relevant to health (e.g. unemployment rates, overcrowding rates, ethnic composition, age and sex structure), local health surveys and any available epidemiological data on incidence and prevalence rates, and morbidity statistics (e.g. cancer registration rates from the national cancer registry, and service use rates in order to assess supply and demand; Stevens 1991).

Some investigators have used action research, and, in particular, rapid appraisal methods to assess the needs of local communities with an emphasis on local people's views and involvement in defining need, priorities and evaluation (Ong *et al.* 1991; Ong and Humphris 1994; Murray and Graham 1995). This involves a collaborative, 'empowering', bottom-up approach to research, using triangulated research methods, for example, community meetings, interviews with key people, postal surveys, feedback of findings to key people and community members and joint development of a plan for action. This requires the use of social science methods to assess needs from a lay perspective, alongside traditional analyses of epidemiological and demographic trends (incidence and prevalence of disease, population trends, mortality patterns) (Fitzpatrick 1994). For example, a health needs assessment in the District of Columbia, USA, used mixed research methods including analysis of routine data, quantitative surveys, qualitative focus groups and assessment of national health areas of priority (Chandra *et al.* 2013). It was reported

that multiple priority areas were identified (asthma, obesity, mental health, sexual health, stress-related disorders) and problems of general access to health services.

The role of epidemiological and demographic research

Epidemiology and demography can provide information on the need for health, though this has to be analysed together with evidence on the effectiveness of health care to be informative about the 'need for health care'. Where the service is of proven benefit (i.e. effectiveness), the demographic and epidemiological data are important *per se* because they are addressing the issue of whether the service is reaching all those who need it (e.g. is cervical cancer screening reaching all women?, are immunisation programmes reaching all children in predefined age groups?). Health services research is the focus for a number of disciplines, each of which plays a complementary role. The diversity of approaches has led to developments in the focus of epidemiological and demographic research as they are influenced by other disciplines and research paradigms. It is impossible to cover the contribution of each discipline to the assessment of needs in one chapter. In Parts 4.2 and 4.3 we concentrate on the main concepts and techniques of analysis within epidemiology and demography.

These disciplines operate within a positivist framework (see Chapter 6). This implies a belief in the scientist as a value-free observer and in the traditional scientific method, in which a hypothesis is generated, and data are gathered and tested objectively in relation to the hypothesis. Within this paradigm, disease in humans is an observable fact, the 'causes' and 'effects' of which are also subject to factual verification under the objective gaze of the investigator. The goal of such an approach is to search for universal explanation, derived from empirical regularities.

4.2 Epidemiology

The role of epidemiology

Traditionally, epidemiology has been concerned with the distribution of, specific causes (aetiology) of, and risk factors for diseases in populations. It is the study of the distribution, determinants and frequency of disease in human populations (Hennekens and Buring 1987). Epidemiology is also concerned with the broader causes of disease. For example, the epidemiological transition model suggests an association between national economic development and health using mortality data. However, this model has been hotly debated, as not all nations fit the model, patterns of mortality within nations change and mortality and health vary within countries by social group. It is also dependent on the way resources are distributed and targeted in societies (Wilkinson 1992).

Mainstream epidemiology examines data on levels of disease and risk factors for disease, while taking environmental factors into account. In contrast, materialist epidemiology is concerned with the role of underlying societal and structural factors. The latter is critical of the reductionist perspective of mainstream epidemiology, which focuses on individual, rather than societal, risk factors (reductionism). The focus on the biological make-up of the individual diminishes the importance of interactions between individuals and, more importantly, the idea that the whole is greater than the sum of its parts is lost. For the exploration of the latter, a more qualitative approach is needed. The limits of epidemiology

can also be found in the way that disease classification is often taken for granted. Although epidemiologists are critical of the difficulties of categorising disease, it is too often assumed that medical classification is a valid research tool, forgetting that diseases, as physical phenomena, can be interpreted in different ways and the act of medical classification itself changes the way we look at and perceive disease. Types of epidemiology, including community, communicable disease, critical, environmental, occupational and social epidemiology, have been described by Moon *et al.* (2000). There is also an increasing focus in epidemiology on 'the life course' (the study of long-term effects on later health, and risk of disease, of physical or social exposures during gestation, childhood, adolescence, young and later adulthood; Ben-Shlomo and Kuh 2002), as well as the health effects of the accumulation of risk, and on genetic epidemiology (Lewis *et al.* 2005).

Epidemiological research

Epidemiological research includes both descriptive and analytical studies. Descriptive studies are concerned with describing the general distribution of diseases in space and time (examples include case series studies and cross-sectional surveys). Analytic studies are concerned with the cause and prevention of disease and are based on comparisons of population groups in relation to their disease status or exposure to disease (examples include case control studies, cohort studies, experimental and other types of intervention studies). However, these distinctions should be interpreted with some flexibility. Rothman (1986) pointed out in relation to epidemiologic research that its division into descriptive and analytic compartments, which either generate (descriptive research) or test (causal) hypotheses (analytic research), is derived from a mechanistic and rigid view of science which is inconsistent with current practice and philosophy. He pointed out that any study can be used to refute a hypothesis, whether descriptive (quantitative or qualitative) or analytic research methods are used.

Causal associations

Epidemiology is faced with difficulties when imputing causality. The difficulties of research design and interpretation of the results include temporal precedence in relation to the direction of cause and effect. This is the confidence that changes in X are followed by subsequent changes in Y, and elimination of the possibility of reverse causation – did depression lead to elderly people becoming housebound or did being housebound lead to depression? (See Chapters 9 and 10.) Experiments deal with reverse causation by the manipulation of the experimental (independent) variable, and measuring the dependent variable usually before and after this manipulation.

Longitudinal survey analyses use multivariable statistics, which can provide estimates for the strength of independent associations over time, where the variables of interest co-vary, and where a spurious association does not exist with other variables, and the hypothetical cause precedes, or occurs simultaneously with, the hypothesised effect in time (e.g. as indicated by the change in the causal variable occurring no later than the associated change in effect). Temporal regression analysis, for example, if carefully designed, can provide information suggesting that the second variable to change did not cause the first, though it is never possible to infer with absolute certainty that the first variable to change caused the second. Other longitudinal analysis techniques include structural equation modelling (e.g. cross-lagged and simultaneous model equations can be specified and estimated). This

technique makes use of the inherent time-ordered data to address causal ordering. These analyses require careful formulation of hypotheses and models of the processes. The strength and duration of reciprocal relationships, and of hypothesised causal effects, are informative. No analysis yields trustworthy inferences about causal structures, due to the reciprocal nature of many potential influences (see Chapters 9 and 10). Difficulties include: chance results; study bias, which may influence the results; intervening variables or bias; and uncontrolled, extraneous variables which can confound the results.

Intervening variable

An intervening variable is an intermediate step in the causal pathway between the independent and dependent variables. In other words, the independent variable (e.g. the experimental or explanatory variable) affects the dependent variable (e.g. the outcome of interest) through the intervening variable. This is also referred to as *indirect* causation. An example is where consumption of fatty food can lead to narrowing of the arteries, which in turn can lead to coronary heart disease, so narrowing of the arteries is the intervening variable.

Confounding variable

A confounding variable is an extraneous factor (a factor *other* than the variables under study), *not controlled for*, which distorts the results. It is *not* an intervening variable (e.g. between exposure and disease). An extraneous factor only confounds when it is associated with the dependent variable (causing variation in it) *and* with the independent variable under investigation. The confounding and independent variables interact together to affect the outcome and their contributions cannot be disentangled. It makes the dependent and independent variables appear connected when their association may be *spurious* (see Figure 4.1).This raises the potential for extraneous variables to confound the results of research, leading to spurious (false) associations and obscuring true effects. If the confounding variable is allowed for, the spurious association disappears.

An example of confounding is where an association is found between cancer and use of hormone replacement therapy. If the cancer is associated with age, then age is a potential confounder; age should be allowed for in analyses (it might simply be that older age is responsible for the association with cancer). Another example relates to the hypothesis that the risk of myocardial infarction is increased among coffee drinkers, compared to non-coffee drinkers. Smoking may be a potential confounding variable in this association, as people who drink coffee also tend to smoke. In Figure 4.2, coffee appears to be associated with myocardial infarction. However, it is believed that smoking is a confounder because it is associated with

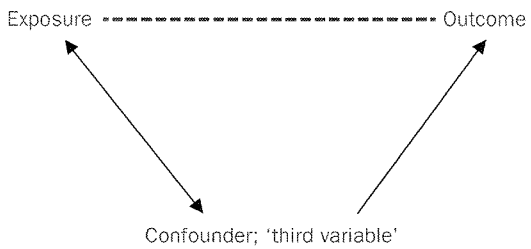

Figure 4.1: Confounders in research

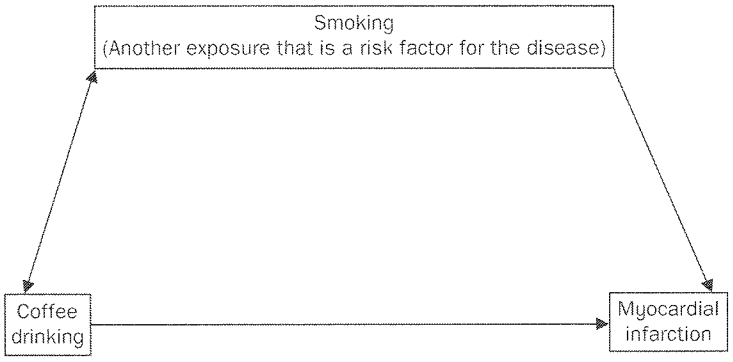

Figure 4.2: Coffee drinking as a confounder in myocardial infarction

both coffee drinking and with myocardial infarction. Smoking status is most likely confounding the association between coffee drinking and myocardial infarction, making it appear that there is a relationship when in fact there is none (see such examples in Varkevisser *et al.* 1991).

In ideal laboratory experiments in natural and biological science, one variable at a time is altered and observed, so that any effects that are observed can be attributed to that variable. This approach is not possible in research on people in their social environment, as human beings differ in many known and unknown ways. Other extraneous variables which are not associated with the independent variable can also lead to misleading results (systematic error or bias; see Chapters 7 and 8).

In epidemiology, the calculation of high relative risks may appear impressive, but important confounding variables may still be missed. A dose–response relationship gives added weight to imputations of causality between variables (e.g. there is a relationship between lung cancer and the number and strength of cigarettes smoked), but this still does not dismiss the possibility of confounding. Confounding is prevented by using randomisation in experimental designs, by restricting the eligibility criteria for entry into studies to a relatively homogeneous group and by matching (see Chapters 10 and 11).

A major research problem is how to decide whether a factor is causally related to an outcome, rather than simply being associated with the factor that is the true causal agent (see Davey Smith and Phillips 1992). This is important because a great deal of research is published which is based on descriptive studies that report associations between the risk of ill health and the exposure to particular factors. Examples include eating beef and risk of Creutzfeldt-Jakob disease, drinking coffee and risk of coronary heart disease, alcohol as a protective factor for coronary heart disease, not being breast fed being associated with low intelligence, use of babies' dummies being associated with low intelligence, use of oral contraceptives and risk of cervical cancer, use of oral contraceptives facilitating HIV transmission and the reverse – use of oral contraceptives protecting against HIV transmission.

It is important to be aware of potential extraneous variables which may confound results at the design stage of the study. These can then be measured and controlled for in the matching process (if used) and/or in the analyses. Age is a common confounding variable and so, to a lesser extent, is sex; most investigators routinely control for age and

sex and analyse results separately for these groups. Randomised, experimental research designs are less likely to suffer from confounding (because of the random allocation to experimental and control groups), particularly if the number of participants is large.

The usual way to address confounding variables is to fit a regression model so that it is possible to examine the relationship between the variable of interest and the outcome while holding other variables constant. This is called 'adjusting' or 'controlling' for other variables. The limitation of this method is *residual confounding*, which arises because of the inadequacy of the measure representing the variable being controlled for (see Glynn 1993). For example, the apparent independent relationship between breast feeding and IQ (i.e. while controlling for social class) may be due to the inadequacy of using the father's occupation as a measure to control for social class effects. Biological plausibility is often appealed to in interpretation of epidemiological associations. However, it is possible to construct plausible mechanisms for many observed effects.

The way forward is for epidemiologists to use triangulated (e.g. multiple) research methods in order to minimise problems of interpretation in the study of the causes and process of disease. Causal arguments are strengthened by similar results being achieved by different studies and by different study designs, in different places and by different investigators. Epidemiologists should also work with social scientists to gather the information that lay people have about their health and lives, the causes of their health and ill health.

Methods of epidemiology

The range of epidemiological methods is described below, and those which are shared across disciplines are described in more detail in later chapters.

Case series and case studies

With the case series method a number (series) of cases with the condition of interest is observed, often using triangulated methods (e.g. questionnaires, data from records, observations) in order to determine whether they share any common features. The observations can be made retrospectively or prospectively, and they are relatively economical in terms of time and resources to carry out. They share the same weaknesses as survey methods, with the additional weakness that the sample is one of cases only, with no point of comparison. However, the method is useful for generating hypotheses. In-depth studies of single cases are known as *case studies*. These are also useful for developing a body of knowledge about a situation and for paving the way for future trials of interventions and cohort studies of conditions (Vandenbroucke 1999). They have a long tradition in medicine and are valuable for describing new diseases and new side-effects of treatment. The case study in relation to qualitative social research methods is described in Chapter 19.

Surveys

Descriptive cross-sectional surveys

Descriptive cross-sectional surveys are based on a representative sample or sub-sample of a population of interest who are questioned at one point in time (see Chapter 9). In

epidemiology, the aim is usually to assess the prevalence of disease, associated factors and associations with service use. For the assessment of prevalence these studies are sometimes conducted in two phases, in which a screening instrument (e.g. a questionnaire measuring depressive symptoms) is used to identify people of potential interest, who are then followed up and assessed in more detail (e.g. with a psychiatric examination to confirm diagnosis). This is sometimes more economical than subjecting the whole sample to a full assessment.

Screening surveys and case finding

Cross-sectional screening and case finding surveys are conducted in relation to the detection of individuals or populations at high risk of disease in order that there can be a health care intervention or health promotion in order to protect them (e.g. as in cardiovascular disease). Population screening surveys have formed the basis for case finding, particularly in surveys of disability and psychiatric problems. Because of the high cost and time-consuming nature of population screens, case finding is now more commonly carried out in opportunistic screening exercises (e.g. detection of cases by questionnaire or record research among people attending a doctor's surgery for any other condition). The problems involved in screening relate to motivating health care professionals and the population to act, as well as ethical issues of invasion of privacy. Such methods are only ethical where the history of the condition is understood, there is a recognisable early symptomatic stage and there is an effective, safe, cost-effective and acceptable treatment available and agreed by policy-makers, clinicians and patients for predefined cases. Screening is generally confined to conditions which are recognised as common and perceived to be important.

Ecological studies

Ecological studies also aim to assess exposure (e.g. 'risk') and disease or mortality. With these, the unit of study is a group of people, rather than the individual (e.g. people in classrooms, hospitals, cities), in relation to the phenomenon of interest. In contrast to individual studies, the unit of analysis is a geographical area or organisation. The groups of interest are sometimes surveyed longitudinally to assess incidence (see Chapter 9). Data collection methods may include questionnaires and record research (e.g. medical records). Individual-level research can miss important area influences (e.g. on health), and hence the unit of analysis in ecological studies is the area level (geographical or organisational). A limitation is the assumption that the characteristics of populations (the study results) are applicable to the individuals within them: the *ecological fallacy* (Martin 2005; Moon *et al.* 2005).

Case control studies

At its most basic, a case control study (also known as a case-referent study) is a descriptive research method which involves comparing the characteristics of the group of interest, such as a group with a specific disease (e.g. ischaemic heart disease), or who have been exposed to a particular risk factor, such as radiation in a nuclear power plant incident (*cases*), with a comparison, or reference, group without the characteristic of interest, or the disease/condition (*controls*). The aim of the comparison is to identify

factors which occur more or less often in the cases in comparison with the controls, in order to indicate the factors which increase or reduce the risk factors for the disease or phenomenon of interest. The analysis will lead to the calculation of an odds ratio, which is an estimate of the contribution of a factor to disease. The number of cases exposed is multiplied by the number of controls unexposed, and this figure is then divided by the product of the number of cases unexposed and the number of controls exposed. It is an approximation to the relative risk, which is a measure of how strongly associated the exposure is with the disease. The extent to which an exposure is more likely to occur in cases than controls is more accurately estimated in longitudinal surveys using relative risk, or rate ratio (the rate of the disease, being the number of cases occurring divided by the population at risk for a period of time). Thus, the case control study primarily aims to investigate cause and effect (see St Leger *et al.* 1992). The starting point is people who have the disease or condition, or who have been exposed to a risk factor. This is in contrast to the epidemiological longitudinal survey which starts with the potential risk factor of interest (e.g. smoking) and then follows up over time people who have and do not have the risk factor, and compares the number of events (e.g. cases of heart disease) in those with and without the risk factor.

In case control studies, then, people can be compared in relation to potentially relevant *existing* characteristics (risk factors) and/or retrospectively in relation to their reported *past* experiences (exposures). Data relating to more than one point in time are generally collected, making case control studies technically longitudinal rather than cross-sectional (longitudinal can relate to more than one point in time in the past – retrospective – as well as to more than one point in time in the future – prospective).

Many textbooks of epidemiology describe case control studies as retrospective. For example, when the group of cases of interest and an unaffected control group have been identified, their risk factors and past exposure to the potential aetiological factors of interest are compared (see Altman 1991). While case control studies are usually retrospective, they can be prospective (high- and low-risk groups are compared in relation to the incidence of disease over the passage of time). Case control studies can also be nested within a descriptive cross-sectional, or longitudinal, prospective study if the latter is sufficiently large to detect significant associations (see Mann 1991; Beaglehole *et al.* 1993). (See Box 4.1.)

Box 4.1 A study of injury amongst steelworkers

An example of a nested case control study is the study of injury in Brazilian steelworkers by Barreto *et al.* (1997). In their cohort study of 21,816 Brazilian steelworkers they found that mortality from motor vehicle injury was twice that in the state population. Therefore they undertook a nested case control study within their cohort to investigate possible socio-demographic, medical and occupational risk factors to explain this increased risk. For each case (all workers who died of motor vehicle injury while employed at the plant during a specific time period), four controls were selected randomly from workers within the cohort who were employed at the time of death of the case, and who were born in the same year as the case. Data for analysis in relation to risk of motor vehicle injury were collected from personnel, industrial hygiene and medical records.

Advantages and disadvantages of case control studies

The main advantages of case control studies are that they are relatively cheap in comparison with experimental designs, they are useful for the study of rarer conditions and they can provide relatively quick results. Case control studies, however, often require large numbers for study, they can suffer from the limitations of potential selection bias among participants, and extraneous, confounding variables (variables that are associated with both the presence of disease and the risk factor or exposure variables) may explain any observed differences between cases and controls.

Adjustment for confounding variables can be made by measuring them and adjusting for them in the data analyses (stratified analysis: the strata are the levels of the confounding variable). However, potential confounders may be unknown and unmeasured. It is also common to use matching techniques (see Chapter 11) in an attempt to limit the effects of extraneous confounding variables, though it is often difficult to match beyond common characteristics (e.g. age and sex). Case control studies suffer from a major limitation in that they are all, in effect, retrospective studies. Even if the cases and controls are followed up over time in order to observe the progress of the condition, the investigator is still starting with the disease or exposure and relating it to past behaviour and events. In particular, the cases may be more anxious to recall past behaviours (i.e. as possible causative agents) in comparison with controls and therefore questionnaire data may be subject to recall, or memory, bias.

A case control study is restricted to recruiting participants from the population *of interest*, and it is important that both groups of participants (cases and controls) should be representative of that population (see Chapter 8 on sample size and sampling). With case control studies, the control group is intended to provide an estimate of exposure to risk in the population from which the cases are drawn, and therefore they should be drawn from the same population – the difference between the two groups being the exposure (the exposed group form the cases). Appropriate controls can be difficult to find. As Altman (1991) has explained, people who do not have the outcome variable of interest may differ in other ways from the cases. It is also common in studies where the cases are hospital patients to use hospital patients with different medical conditions as controls. This can lead to bias because the conditions the controls are suffering from may also be influenced by the variable of interest, leading to underestimates of the effect of that variable in the cases (e.g. smoking is associated with several conditions). Ebrahim (1990) has described these problems, and pointed out that many investigators of hospital-based cases (e.g. stroke) now use two sources of controls: a community group (e.g. drawn from general practitioners' lists or local population register) and a hospital control group. As the rule for selecting controls is to ask whether cases and controls are equally likely to have been exposed to the variable of interest, then any doubt about this implies that the comparison of cases with controls may be biased.

One risk of the rigorous matching of multiple variables in case control studies is the 'controlling out' of the variable of interest. This is referred to as *over-matching*. This means having a control group that is so closely matched (and therefore similar) to the case group that the 'exposure distributions' differ very little. Rothman (1986) argued that this interpretation of over-matching is based on a faulty analysis which fails to correct for confounding variables – and is corrected if stratification by the matching factors is used in the analysis. He argued that the 'modern interpretation' of over-matching relates to 'study efficiency rather than validity'.

Research using documents

Epidemiologists use official statistics (which they call 'vital' statistics) on mortality (displayed by socio-demographic factors, area mortality and occupational mortality) and morbidity (e.g. on cancer registrations, congenital malformations and infectious disease surveillance). Their use plays a central role in disease surveillance. There are many problems with the use of official statistics because diagnostic criteria and disease classifications may change over time, and diagnostic definitions may also vary by area, making comparisons difficult. While data such as birth and death registrations are complete in the 'developed' world because it is a statutory duty to register them, other data may not be (e.g. routine patient administration data reporting types of procedures performed and disease classifications of patients discharged).

Prospective, longitudinal cohort surveys

These 'follow-up' studies are intended to assess the incidence of disease and the potential causative agents of disease in a population which divides itself 'naturally' into exposed and unexposed groups. The term 'natural' refers to the fact that they are not artificially manipulated by the research design as in experimental studies. There are two types of longitudinal study: panel and trend (see Chapter 9). With panel surveys there is no turnover of membership. However, account also needs to be taken of the time over which the survey members were observed (as well as the size of the population). With the fixed population in the panel survey, the population gradually diminishes in size as its members die and cease to be at risk of becoming 'a case'. Thus epidemiologists often use longitudinal trend surveys which are composed of dynamic populations (i.e. there is turnover of membership). 'Cohort' means the sample shares a common factor (e.g. age).

Life course approaches

Social science has long supported the analysis of processes that operate across a person's life in order to understand later outcomes of interest. Sociology of the life course is based on the assumptions that people's lives are embedded in, and shaped by, historical context; individuals construct their own lives through their choices and actions, and within the limitations of social and historical constraints; lives are intertwined via social relationships; and the meaning and impact of a life transition depend on when it occurs (see Clausen 1986). Data are collected from multiple sources, including longitudinal surveys, censuses, and life history interviews. The development of such life course approaches in epidemiology emerged later during the 1990s (Blane *et al.* 2007). Kuh *et al.* (2003, p .778) defined life course epidemiology as

> the study of long term effects on later health or disease risk of physical or social exposures during gestation, childhood, adolescence, young adulthood and later adult life. The aim is to elucidate biological, behavioural, and psychosocial processes that operate across an individual's life course, or across generations, to influence the development of disease risk.

A classic example of the value of life course approaches is analysis of a birth cohort from 1946 which demonstrated the importance of childhood illness (in turn influenced by parents' social class) to health in adult life (Wadsworth 1986).

Life course epidemiology involves building and testing theoretical models of pathways that link exposures across the life course to specific outcomes. The approach involves analysis of temporal ordering of exposures and their inter-relationships, and is complex, requiring time-related study designs. The ideal method uses birth cohort studies, in which data is collected from birth and throughout life. In practice, as with all longitudinal designs, the surveys rely on retrospective questioning at each follow-up phase, and are subject to recall bias.

The randomised controlled trial (RCT)

This is the ideal, true experimental method for the evaluation of the effectiveness of health services and interventions in relation to specific conditions. The method involves two or more groups who are treated differently, and random assignment to these groups. These features require the investigator to have control over the experimental treatment and over the process of random assignment between groups (see Chapters 10 and 11).

The natural experiment

At a basic level, the experiment is a situation in which the independent (experimental) variable is manipulated by the investigator or by natural occurrence. An investigation in a situation in which the experimental setting has been created naturally is known as the *natural experiment*. The classic and most popular example is John Snow's study of cholera in London in 1854, which established the foundations of modern epidemiology as a form of systematic analysis. (See Box 4.2.)

Box 4.2 Snow's study of cholera in London

At the time of the 1848 cholera outbreak in London several water companies supplied piped drinking water. Snow (1860) compared the mortality rates from cholera for the residents subscribing to two of the companies, one of which piped water from the River Thames near the point where large amounts of sewage were discharged, and the other which piped water from a point free of sewage. In effect, the natural experiment permitted Snow to obtain data on around 300,000 people, who spanned all socio-demographic groups, and who were divided naturally into two groups without their choice: one receiving water containing sewage and the other receiving water free from impurity. Snow used a map and plotted the location of the outbreak, having already noted the cases to be clustered around Soho in London. Snow discovered that people who had drunk water from the pump in Broad Street (now called Broadwick Street), supplied by the company drawing its water from the contaminated part of the Thames, were more likely to contract cholera than those who had not. Snow arranged for the removal of the handle to the pump and the outbreak stopped (though it had apparently already peaked). This is also a good example of how epidemiology is concerned with populations rather than individuals (see Lilienfeld 2000; Sandler 2000; Vandenbroucke 2000; Medical Research Council 2011).

Natural experiments worldwide include examination of whole population suicide rates in Sri Lanka in relation to the introduction of legal restrictions on pesticide imports (Gunnell *et al.* 2007); analysis of the use of health facilities to give birth, infant and child mortality in India, in relation to cash incentives to use health facilities to give birth (Lim *et al.* 2010);

and analysis of all-cause and cause-specific mortality in the population of Hong Kong in relation to the introduction of legislation to restrict the sulphur content in fuel (Hedley *et al*. 2002) (see Medical Research Council 2011).

Field experiments

Field experiments, or trials, are research studies in a natural setting in which one or more independent variables are manipulated by the investigator, under situations as controlled as possible within the setting. Field trials usually involve the study of healthy individuals in relation to the health outcome of preventive measures aimed at individuals (e.g. supplementation of diet with vitamins). With this method, large numbers of people have to be recruited in order to obtain an adequate proportion of them who will go on to contract the disease having received the intervention. This makes the method expensive. The difficulties of controlling intrinsic and extrinsic factors are also greater than in tightly controlled laboratory or clinical settings.

The *true* experiment, with the randomisation of participants to intervention or control group, and with pre- and post-testing, is the ideal model for this (see Chapter 10 for the distinction between the basic and the true experimental method). However, in practice, random allocation to the intervention is not generally feasible. Results are more difficult to interpret without random allocation of people to exposed and non-exposed (control) groups because of the potential for unknown extraneous variables which may confound the results (see Chapters 10 and 11).

Community intervention experiments

Community intervention experiments, or trials, involve a community-wide intervention on a collective (rather than individual) basis (e.g. in order to study the health outcome of water fluoridation, which is aimed at communities and not allocated to individuals). With this method, entire communities are selected and the exposure (e.g. the fluoridation) is assigned on a community basis. The community is defined either as a geographical community or as units in social groupings (e.g. hospital wards, school classrooms). Ideally, the *true* experimental method is adhered to, and the assignation of communities to the exposure or no exposure group is carried out randomly. With large numbers of people involved, this is rarely feasible and geographical comparisons are frequently made between areas exposed and not exposed (without randomisation), and the effects of the exposure. If there are no differences between the communities in their socio-demographic or other relevant characteristics, this non-random element may have little effect. Again, results are more difficult to interpret without random allocation of people to exposed and non-exposed groups (see Chapters 10 and 11). There can also be problems with sample size. An intervention community is commonly compared with one control community. This is a weak design, which is equivalent to a clinical trial with one patient in each treatment group, and no information can be provided on variation between communities (Hays and Bennett 1999).

Assessing morbidity, mortality, incidence and prevalence

Morbidity and mortality

The ideal first step when assessing the need for health care is the epidemiological survey of a defined population to establish the incidence (number of new cases) and

prevalence (all existing cases) of morbidity in relation to the disease or condition of interest. Mortality patterns also require analysis. While figures on mortality by cause and by socio-demographic characteristics are available from official sources in the developed world, data on morbidity patterns (apart from cancer) are not routinely collected. In Britain, with a nationalised health service, some data are available centrally. These are collected from NHS hospitals, and cover numbers of patients discharged with their standard disease and operation coding. However, these data may be incomplete and subject to coding errors, and only represent people who are admitted to hospital (and who form the tip of the iceberg of illness in the community). Surveys of morbidity reported in general practice and comprehensive community health surveys are only carried out on an ad hoc basis. However, as noted earlier, it is sometimes possible to apply their findings to other populations if they are similar in structure. Except in relation to conditions where case-fatality is high and constant over time, and where the length of time with the condition is relatively short (e.g. as in some cancers), mortality statistics cannot be used as proxies for morbidity.

Information will also be required on the severity of disease and on current treatment patterns (in order to calculate the size of the gap between estimated need for a service and the expressed and satisfied demand for it), survival time and mortality rates. All this needs to be collected and analysed by age, sex, socio-economic group and ethnic status at minimum (where relevant), and an estimate should be made of the proportion of the population at risk and increased risk of the disease/condition. This requires precise definitions of the condition, rigorous assessments of health status in relation to the condition and agreement on clear and correct cutoff points for effective treatment (e.g. the level of high blood pressure which can be effectively treated). The last is essential in order to calculate the number of people who are likely to benefit from the service.

Incidence

Incident cases are new instances (of disease or death) which occur in a defined *time period*. *Incidence* refers to the number of new cases in a population in a defined time period. The *cumulative incidence* rate is the number of cases (the numerator) that occur (rate of occurrence) in a defined time period divided by the number of people in the population (the denominator) at the beginning of the period. It is more common to calculate the incidence rate of a disease over a specific period of time (e.g. a year); this is the number of *new* cases of the disease over the time period divided by the number in the population at risk (more specifically, the total time each member of the population remained at risk). Incidence is usually expressed as a percentage, or as number of cases per 1000 or per 100,000 people in the population.

Prevalence

The *prevalence* of a disease at a *specific point in time* is calculated by taking the *total number* of existing cases of the disease at that time divided by the number in the population at risk. With *point prevalence* (the number of cases at a certain point in time) a very short time period is examined (e.g. days or a few weeks). With *period prevalence* (the number of cases during a specified period of time) a longer time period is examined (e.g. weeks or months). *Lifetime prevalence* is measured by taking the number of people who have had the condition/disease at least once during their lifetime. Prevalence is usually expressed in terms of the number of cases (e.g. of disease) in a population at one point

in time per 1000 or 100,000 population. The formulae for the calculation of incidence and prevalence ratios can be found in Rothman (1986).

Person time at risk

The person time at risk is the length of time each individual has been under observation without developing the disease. For a group of four people, one of whom was lost to follow-up after one year, one of whom developed the disease after two years and two of whom were still free of the disease after four years, the total person time at risk would be 11 years. Direct measures of the length of time a person is at risk are not available from routine ('vital') official statistics on mortality. Instead, the population at the mid-point of the time period of interest, multiplied by the length of the period (e.g. in years), is taken as an estimate of the person time at risk.

Case-fatality

This is a form of cumulative incidence and is related to the survival rate of the disease of interest. It measures the proportion of people with the disease who die within a defined period of diagnosis.

Odds ratio

While one way of comparing two groups (e.g. cases and controls) in relation to the disease of interest is to calculate the ratio of the proportions of those with the disease in the two groups, another method is to calculate the *odds ratio*: the ratio of the odds (loosely, a type of probability) of the disease ('event') in the two groups. This is an estimate of the contribution of a factor to disease. The calculation of odds has been clearly explained by Deeks (1996). The odds are calculated as the number of events divided by the number of non-events. More precisely, the number of cases exposed is multiplied by the number of controls unexposed. This figure is then divided by the product of the number of cases unexposed and the number of controls exposed. It is an approximation to the relative risk, which is a measure of how strongly associated the exposure is with the disease.

If the odds of an event are greater than 1, then the event is more likely to occur than not. If the odds are less than 1, the chances are that the event will not occur. The odds ratio is calculated by dividing the odds in the treated or exposed group by the odds in the control group. Epidemiologists attempt to identify factors that cause harm with an odds ratio of greater than 1. Clinical studies investigate treatments which reduce event rates, and which have an odds ratio of less than 1. The odds ratio can be used as an approximation of the relative risk in a case control study.

Measures of effect

In epidemiological terms, effect refers to the difference in disease occurrence between two groups of people who differ in relation to their exposure to the causal agent. There are three types of effect: *absolute effects* (differences in incidence, cumulative incidence or prevalence), *relative effects* (the ratio of the absolute effect to a baseline rate), and *attributable proportion* (the proportion of the diseased population for which the exposure to the causal characteristic was one of the causes of that disease). Measures of effect include relative risk, attributable risk and population attributable risk.

Relative risk

The relative risk, or rate ratio, is the incidence rate for the disease in the population exposed to a phenomenon relative to (divided by) the incidence rate of disease in the non-exposed population.

In other words, the relative risk indicates how much more likely a given disease or event is in one group compared with another. The relative risks of disease (e.g. lung cancer) in relation to the phenomenon under investigation (e.g. smoking) can be directly calculated if longitudinal survey methods are used, because the incidence and prevalence of the condition in the (exposed and unexposed) study population are known. It is also possible to calculate *confidence intervals* for relative risks. In a case control study with a sample of cases and a sample of controls, it is only possible to estimate relative risks indirectly (in the *odds ratio*). Only estimation is possible because a case control study does not include a sample of exposed and unexposed members (just a sample of cases and a sample of controls), and therefore the prevalence of disease is unknown.

Attributable risk

The attributable risk relates to the absolute effect of the exposure and is the difference between the incident rate in the exposed population and the incident rate in the non-exposed population. In other words, attributable risk indicates on an absolute scale how much greater the frequency of the disease or event is in one group compared with the other. This is an absolute measure of risk which is suited to the analysis of individuals, and not generalisable.

Population attributable risk

This gives a measure of the excess rate of disease in the whole population that can be attributed to the exposure of interest. It is calculated by multiplying the individual attributable risk by the proportion of exposed individuals in the population. It measures the population burden (*need*). The data are not generalisable.

Numbers needed to treat

Numbers needed to treat measures how many people need to receive the intervention (e.g. prescribed medication) for a given period in order that one more person will have the specified successful outcome, compared with the number who would have that outcome without the intervention. This is a meaningful way of expressing the benefit of the intervention. In a trial the number needed to treat is the inverse of the difference between the proportion of events in the control group and the proportion of events in the intervention group. An alternative model for the number needed to treat has been put forward as the inverse of the proportion of events in the control group multiplied by the reduction in relative risk (Chatellier *et al.* 1996). Rembold (1998) has proposed a formula for numbers needed to screen for use in evaluations of the efficacy of disease screening.

Comparisons of rates and standardisation

The comparison of rates across different populations can be misleading and therefore the standardisation of rates is essential in order to reduce any distortions. These methods are discussed, with demography, in Part 4.3.

4.3 The role of demography

Pure demography is the study of populations in terms of the numbers of people, and population dynamics in relation to fertility, mortality and migration; the broader area of population studies addresses the issues of *why* observed changes occur, and the consequences of these (Grundy 1996).

Changes in population structures are the result of changes over time in fertility, mortality and, to a lesser extent, international migration. Historically most countries had high levels of fertility and mortality. As major infectious diseases were controlled and declined, overall mortality levels declined and life expectancy at birth increased, while fertility remained high. One consequence was reduced infant mortality and a high percentage of children and young adults because younger age cohorts increase relative to older age cohorts. Populations begin to age when fertility falls and mortality rates continue to improve or remain low. Successive birth cohorts may become smaller. Countries that have low fertility and low mortality have completed what demographers call the 'demographic transition'. The term 'epidemiological transition' is used to describe the transition from relatively high to low mortality patterns, associated with changes in mortality by age and sex (Omran 1971); and the term 'health transition' refers to changes in the response of societies to health and disease (see Grundy 1996).

Demographical methods in relation to assessing need

The understanding of how populations change is vital to the assessment of needs for health services in order to plan services accurately (e.g. number of maternity beds and long-stay care places for elderly people that will be required). Demographic and social data (known as 'socio-demographic data') by definition provide information on the social and demographic characteristics of populations, and on areas of social deprivation. This information can be analysed in relation to mortality patterns, any existing data on morbidity for the populations of interest and service allocation. Such data have implications for 'need for health', though they cannot provide information on needs for effective health services.

Grundy (1996) has described how demography requires information about population 'stock' and 'flows' in and out of the population. The traditional demographic sources are population censuses and vital registration systems, supplemented with data from population surveys. National socio-demographic data are collected using the census, and local population data are derived from this. Interim profiles use the last census as the baseline and make adjustments (population estimates or informed guesses) for changes in the population since the last census was conducted. At the local level some further adjustments might be made in the light of local information. National data are available on births, marriages and deaths in populations, and also cancer registrations, as these are registered events. Similarly, information on immigrations and emigrations is available. From the information contained in the registrations it is possible to compile national figures on, for example, age and sex in relation to births and deaths. A wide range of analyses are carried out in relation to mortality (e.g. cause of death using International Classification of Disease codes, area, age, sex, socio-economic group, marital status). In Britain these analyses are carried out and published by the Office for National Statistics

(formerly the Office of Population Censuses and Surveys). There are potential sources of bias and error in each of these sources. For example, certain sub-groups of the population may not be included in censuses (e.g. students, people temporarily away from home); there may be under-reporting of age in censuses; the cause of death recorded on death certificates may reflect changing knowledge or the training and perspective of the certifying doctor.

Using knowledge about current population structures, together with assumptions about future fertility, mortality and migration patterns, demographers can make predictions about future population structures. The method used for calculating population projections (estimates of future population numbers and socio-demographic characteristics, e.g. age and sex) is known as the *demographic component method*. Starting with a base (e.g. the census), assumptions are made about future birth, death and migration rates. Death rates are easier to predict than birth and migration rates as the latter can both be affected by economic, political and social circumstances. The range of demographic concepts, techniques, problems and methods of calculation has been described by Grundy (1996).

Rates: births and deaths

Population growth

This is a function of the balance of births and deaths, taking into account the extent of net migration. A common indicator of growth is the crude rate of natural increase (the difference between the crude birth rate and the crude death rate), taking migration into account (see Grundy 1996, for further details).

Crude birth rates

The crude birth rate is the number of births in a particular year divided by the total in the population and, at its simplest, multiplied by 100 (to express as a percentage). However, it is more usual to express birth and death rates per 1000 people in the population, and the multiplication is by 1000 instead.

Specific birth rates

Because it can be misleading to compare populations in relation to their crude birth rates (e.g. some populations may have higher proportions of males, which might explain their lower birth rates), it is necessary to use an estimate of the number of women of childbearing age in order to calculate the *general fertility rate*. This is calculated by the number of births divided by the number of women of childbearing age, multiplied by 1000.

Crude death rates

The crude death rate is the number of deaths in the population, expressed, for example, per 1000 total population. This is usually calculated, in relation to a particular year, by the number of deaths that year divided by the total population that year, multiplied by 1000.

It can be misleading to compare crude death rates of populations because they may have different age structures. For example, a country or geographical area may have a higher proportion of deaths (crude death rate) simply because it has more elderly people living in

it or more males (and males have a shorter life expectancy than females). Therefore, it is essential to calculate *age-specific death rates for each sex* before comparisons can be made.

Age-specific death rates

The age-specific death rate is usually presented as so many deaths per 100,000 male or female population in the age group of interest per year. In relation to either males or females in a specific age group, for a particular year, the calculation is the number of men or women in a particular age group (e.g. 65–69 inclusive) dying that year, divided by all men or women in that age group, multiplied by 100,000.

Life expectancy

Age-specific death rates have the disadvantage of providing several figures for analyses, rather than just one. Therefore demographers and epidemiologists prefer to calculate and analyse life expectancy and standardised mortality ratios.

Life expectancy is a measure of the average (mean) length of life. Because the average length of life is affected by death rates in many different years, life expectancy is calculated from the average lifetime of a hypothetical group of people. This is based on the assumption that the age-specific death rates in the population of interest in a particular year would continue unchanged in all subsequent years. This allows hypothetical average life expectancy to be calculated and defined as the expectation of life at birth for a population born in a specific year. Although it differs from actual life expectancy in relation to individuals, because the latter do change over time, it does dispense with the requirement to wait until everyone who was born in a particular year has died before life expectancy rates can be calculated.

Standardisation

If the incidence or prevalence of disease or mortality is to be compared between populations, then it is necessary to ensure that the crude rates are calculated from data which are complete and accurate and not misleading. Crude rates are misleading. In theory, the age-specific rates should be compared, but it is cumbersome to deal with a large number of rates. The alternative is to calculate a single figure. In order to be reliable, the single figure must take account of different population structures. This is known as a standardised rate. For example, the standardised mortality rate refers to deaths per 1000 of the population, standardised for age.

Although it is common to standardise by age, and it is possible to analyse males and females separately, there are many other variables which are associated with mortality and morbidity in a population which are not taken into account (e.g. ethnic origin, socio-economic status). Thus analyses must always be interpreted with caution.

The two common methods of calculating standardised rates are direct standardisation and indirect standardisation. The indirect method is generally used. If sample sizes in the index population (population of interest in the area of interest) are small, there can be an increase in *precision* over the direct method, and the direct method can only be applied if the distribution of cases (of morbidity) or deaths in the index population is known. As these distributions are often unknown, the indirect method is generally used, though the direct method of standardisation is generally more consistent if sample sizes in the index population are large enough.

Direct standardisation

The direct method of standardisation has the advantage that it is relatively straightforward and likely to be more consistent than indirect standardisation. If one index population is to be compared with another, it is possible to take the ratio of the two directly standardised rates to yield the comparative incidence index or comparative mortality index. However, the sample sizes in the index population have to be sufficiently large, and the distribution of cases or deaths in the index population needs to be known for this method.

In order to overcome the problem of differences in the structures (e.g. age) of the populations to be compared, a standard population is selected (it may or may not be one of those under study), and the age-specific (or other relevant characteristic) rates (morbidity or mortality) of the index population are applied to the standard population. This provides the number of cases in each age group that would be expected if the index population rates applied in the standard population. The expected number of cases across the age groups is totalled to obtain the total number of expected cases. The standardised incidence *rate* for the index population is the total of these expected cases across the age groups, divided by the total in the standard population.

Indirect standardisation

Indirect methods of standardisation are often preferred because, unlike the direct method, the indirect method does not require knowledge of the age-specific rates in the index population and because the numbers of cases at each age may be small, and thus the age-specific rates of the index population used in the direct method may be subject to considerable sampling error.

The 'standardised incidence ratio' for morbidity and the 'standardised mortality ratio' for the study of mortality are derived using indirect methods of standardisation. The steps for the calculation of each are identical, except that the former is based on a set of standard age-specific incidence rates and the latter is based on a set of age-specific mortality rates (total or for the cause of death of interest).

Standardised incidence ratio

With the indirect method of standardisation for both incidence and mortality, a standard set of age-specific rates in relation to the variable of interest needs to be obtained (e.g. age-specific rates for breast cancer in the total population of females). These standard rates are applied to the index population (the predefined population in the area of interest) in order to determine the number of cases expected in each age group in the index population, on the assumption that the index population experiences incidence of the variable under investigation at the standard rates. These expected cases in the index population are totalled over the age groups to obtain the total number of expected cases in the index population. The total of the observed index cases is divided by the total number expected in order to obtain the standardised incidence ratio. The crude rate in the standard population is multiplied by the standardised incidence ratio to give the standardised incidence rate in the index population.

Standardised mortality ratio

In relation to mortality, the steps are the same as for the standardised incidence ratio (except that mortality, not disease incidence, is the variable of interest), and the ratio is called the standardised mortality ratio (SMR).

The SMR compares the standard mortality rate for the standard (whole) population with that of particular regions or groups (index population), and expresses this as a ratio. The standardised rate in the index population is obtained by multiplying the crude rate in the standard population by the SMR. The procedure for the calculation of the SMR is explained further below.

SMRs are a method of indirect standardisation and are calculated in order to be able to make comparisons of death rates from all causes and mortality from a single cause between geographical areas. They can be calculated for both sexes combined or for just males or females. For the SMR, the crude death rates for particular diseases are calculated (see earlier), often separately for each sex. In order to avoid using small numbers it is more usual to calculate crude death rates from specific causes per 100,000, or per 1,000,000, rather than per 1000. However, the age structure of the population must also be taken into account. As was previously pointed out, this can be done by calculating the age-specific death rates for the disease of interest for each index area and comparing them, though this has the disadvantage of providing several figures (for each age group). The alternative is to use *age standardisation*.

For age standardisation a standard population is selected as a reference point for the geographical area of interest (e.g. the population of a whole country). The SMR is then calculated by using the age-specific death rates for the standard population. A clear example of this has been provided by McConway (1994a):

> So to work out the SMR for male deaths from lung cancer in West Yorkshire, using England and Wales as the standard, the first step would be to find out the age-specific death rates for lung cancer for men in England and Wales. These can be used to work out how many men would have died of lung cancer in West Yorkshire if the impact of the disease on men of any given age there was the same as it was nationally.

The SMR for deaths from a particular disease is then calculated by expressing the actual number of deaths in the group of interest (e.g. number of female deaths from breast cancer) in the index area (geographical area of interest) as a percentage of the expected number of deaths from the standard population data. For example, if the actual number of female deaths from breast cancer in the index population (in the geographical area of interest in England) was 800, and if the application of national female breast cancer rates to the index population (in the geographical area of interest) yielded an expected figure of 700, then the SMR is calculated by expressing the actual number of deaths (800) as a percentage of the expected number of deaths (700). This gives an SMR of 114, and as this is over 100 it means that 14 per cent more females died of breast cancer in that area than would have been expected from national figures, allowing for differences in age structure. It is better to consider the upper and lower confidence limits for an SMR, as these tell us whether the mortality differs significantly from the national average.

Analyses of survival

Survival analysis and life tables

Survival analyses, leading to the estimation of survival rates (e.g. a five-year survival rate), can be carried out in relation to the period of time between a specific event (e.g. medical diagnosis) and death or in relation to a range of other *end-points* of interest (e.g. in relation to onset or diagnosis, recurrence of condition, readmission to hospital,

success of therapy, and so on; or, in relation to marriage, divorce or widow(er)hood). The method of calculation and the formulae for the construction of life tables have been described by Bland (1995). Grundy (1996) has described the concept of *life tables*. Life tables are derived from age-specific mortality rates and show the probability of dying, and surviving, between specified ages. They permit life expectancy and various population projections to be calculated. To carry out the calculation for survival times for people with a specific cancer, for example, the investigator needs to set out, for each year, the number of people alive at the start, the number who withdrew during the year, the number at risk and the number who died. For each year, the probability of dying in that year for patients who have reached the beginning of it is calculated, and then the probability of surviving into the next year. Then the cumulative survival probability is calculated: for the first year this is the probability of surviving that year; for the third year it is the probability of surviving up to the start of the third year and so on. From this life table, the survival rate (e.g. five-year survival rate) can be estimated (Bland 1995).

Mortality compression

Where infant mortality is high but declining, as in developing countries, most of the improvements in life expectancy at birth result from the survival of infants. Once infant and child mortality are low, as in the developed world, the gains in life expectancy are greatest among the oldest members of the population. As mortality rates among elderly people decline, more people survive to older ages. Most of the common health problems in old age are chronic, rather than immediately life-threatening. There is evidence that physiological functioning is declining more slowly with age than was previously thought, though it appears that women can expect to spend more of their years in a disabled state than men, negating some of the benefits of longer life expectancy among females (Manton 1992; Kinsella 1996). With these trends (or epidemiological transitions), conventional indicators of the health of the population (e.g. life expectancy) are less useful. Thus, research in demography is also focusing not simply on the loss of healthy life years due to disability (e.g. the disability-adjusted life year), but on whether morbidity and functional disability in old age are compressed into a shorter and later time period than previously or whether it spans the whole range of later years (i.e. healthy life expectancy, often termed active life expectancy, quality-adjusted life expectancy and disability-free life expectancy).

Disability-free life expectancy (DFLE)

This is an indicator that aggregates mortality and morbidity data for a population into a single index (Sullivan 1971; European Concerted Action on Harmonization of Health Expectancy Calculation in Europe 1996). It represents the average number of years that a person of a given age may expect to live free of disability (Colvez 1996). Demographers have used a range of different survey and mortality tables for their calculations of DFLE (Jitapunkel *et al.* 2003), which creates difficulties in making comparisons across the world (Robine 1992).

The calculation of DFLE requires the availability of standard, current mortality tables (life tables), and data on the prevalence and incidence of morbidity from representative longitudinal survey data with valid and reliable measures of disability. However,

longitudinal data on incidence are less often available and most investigators use data from cross-sectional surveys in their formulae. Calculation of DFLEs is usually based on the method of Sullivan (1971). With this method, a standard cross-sectional life table is taken which gives the number of person years between two ages. This is subdivided using cross-sectional survey data on age-related prevalence of permanent and temporary disability into years with and without disability. A new life expectancy is then calculated using only the years lived without disability. Thus, the rate of permanent and temporary disability is used to estimate the number of years free from disability: 'For example, if 1,000 person-years are lived between ages 75 and 79, and 30 per cent of the population aged 75–79 years suffer from disability, then the number of years free from disability is said to be 700' (Bisig *et al.* 1992).

This method, using cross-sectional data, is inevitably crude. In particular, the level of DFLE is influenced by the measures of disability used in the studies taken for the calculations. Further, as Colvez (1996) pointed out, data on the prevalence of disabilities derived from a series of cross-sectional surveys are not able to provide information on incidence or the probabilities of becoming disabled the *next* year. Cross-sectional surveys can only provide population profiles for a defined time period, and they cannot provide data showing the turnover of people from one category of health status to another.

Sullivan's (1971) method has been criticised by Newman (1988) and Péron (1992), as it does not take into account the reversibility of disabled states. Péron suggests that the correct method is to construct a table showing transitions into and out of states of disability and good health. This presupposes knowledge of the rates of transition from good health to disability and vice versa, and of the mortality rates of disabled and non-disabled people for the same period. Ideally, this requires robust and representative, systematically collected longitudinal survey data on disability, which are rarely available.

Developments include extending the method of potential gains in life expectancy to DFLE (Colvez and Blanchet 1983; Colvez 1996). The potential gain in life expectancy owing to the elimination of all deaths from specific causes is added to the potential gain in DFLE owing to eliminating disabilities due to the same cause.

Disability-adjusted life years (DALYs)

The World Bank (1993) adopted a slightly different approach with the development of DALYs. DALYs estimate the loss of healthy life using international mortality data. With this procedure, the number of years of life lost was estimated for each recorded death in 1990. This was then taken as the difference between actual age at death and the life expectancy at birth which would have characterised a country with a low mortality rate. The loss of healthy life owing to disability was estimated using information from morbidity surveys or expert opinion, and the typical duration of each disease was combined with a weighting to reflect its likely severity. Finally, death and disability losses of healthy life were combined to give the number of years of healthy life lost owing to death or disability (see Curtis and Taket 1996).

Potential years of life lost (PYLL)

The PYLL compares the life expectancy of the whole population with that of particular groups or geographical areas, and expresses it as a ratio.

Summary of main points

- Dictionary definitions of need focus on 'want', 'require' and 'necessity'. The definition of health needs varies between academic disciplines.
- Health policy-makers base health need on a disease model and define it in relation to the need for effective health care and preventive services.
- Lay knowledge is vital for the understanding of health and health care needs.
- The methods of epidemiology and demography can provide information on the need for health; this has to be analysed with other data on the effectiveness of health care to be informative on the need for health services.
- Epidemiology is concerned with the distribution of, causes of and risk factors for diseases in populations.
- Demography is the study of populations in terms of the numbers of people, and population dynamics in relation to fertility, mortality and migration. Population studies, a broader area of demography, addresses the issues of why changes occur and their consequences.

Key questions

1 Define the concept of need.
2 Distinguish between need for health and need for health services.
3 What are the ideal steps in the assessment of the need for health services?
4 What are the main research methods used by epidemiologists?
5 Define a confounding variable.
6 Explain the concept of over-matching in case control studies.
7 Distinguish between incidence and prevalence.

Key terms

attributable risk
case control study
case finding
case series study
cohort
community intervention experiments
confounding
cross-sectional study
demand
demography
disability-free life expectancy (DFLE)
ecological study
effect
epidemiology
extraneous variable
field experiments
health need
healthy life expectancy
incidence
intervening variable
life expectancy
life tables
mortality compression
natural experiment
need
needs assessment
population attributable risk
prevalence
prospective (longitudinal) survey
randomised controlled trial (RCT)

rate
ratio
relative risk
screening surveys
spurious association
standardisation
standardised incidence ratio
standardised mortality ratio
survival analysis

Recommended reading

Bland, M. (1995) *An Introduction to Medical Statistics.* Oxford: Oxford Medical Statistics.

Carneiro, I. and Howard, N. (2011) *Introduction to Epidemiology.* Maidenhead: Open University Press.

Deeks, J. (1996) What is an odds ratio? *Bandolier, Evidence-based Health Care*, 3: 6–7.

Grundy, E. (1996) Populations and population dynamics, in R. Detels, W. Holland, J. McEwan and G.S. Omenn (eds) *Oxford Textbook of Public Health*. Oxford: Oxford University Press.

Martin, R.A. (2005) Epidemiological study designs for health research and evaluation, in A. Bowling and S. Ebrahim (eds) *Handbook of Health Research Methods: Investigation, Measurement and Analysis.* Maidenhead: Open University Press.

Moon, G., Subramanian, S.V., Jones, K. *et al.* (2005) Area-based studies and the evaluation of multilevel influences on health outcomes, in A. Bowling and S. Ebrahim (eds) *Handbook of Health Research Methods: Investigation, Measurement and Analysis.* Maidenhead: Open University Press.

St Leger, A.S., Schnieden, H. and Wadsworth-Bell, J.P. (1992) *Evaluating Health Services' Effectiveness*. Buckingham: Open University Press.